ChinaXiv合作期刊
图书情报工作
第 65 卷 第 9 期 2021 年 5 月

多特征融合的关键词语义功能识别研究*

■ 张国标[1,2] 李鹏程[1,2] 陆伟[1,2] 程齐凯[1,2]

[1] 武汉大学信息管理学院 武汉 430072 [2] 武汉大学信息检索与知识挖掘研究所 武汉 430072

摘 要:[目的/意义]关键词作为一种能够揭示学术文本主题及核心内容的词汇或术语,对其进行功能识别可为知识和文献的快速、精确获取提供底层索引支持。[方法/过程]针对现有研究在关键词上下文建模中多局限于文本层面的符号语义表征,在深入挖掘文献行书规律的基础上,提出一种基于多特征融合的词汇功能识别模型。模型在采用BERT模型捕获关键词上下文依赖特征的同时,融合关键词在关键词列表和全文中的位置信息以及词汇功能先验知识信息,继而采用注意力机制和前馈神经网络对关键词进行问题方法的语义功能判别。[结果/结论]实验结果显示,关键词的位置信息和先验知识均能有效提升关键词语义功能识别性能,其中先验知识对识别效果的提升有较大贡献。

关键词:词汇功能识别 学术文本 关键词 BERT 多特征融合

分类号:G250

DOI:10.13266/j.issn.0252-3116.2021.09.010

1 引言

关键词作为学术文献中映射全文主题内容的词汇或术语,是一种能够对文本内容和主题高度凝练概括的功能性词汇,亦可为信息检索、知识组织以及大规模文本计算提供多层次的语义标签。然而,过于着重精简性的关键词牺牲了大量的上下文信息,由此造成的语义功能模糊、使用意图不明使其在脱离原文后便难以释读。相对于其他的检索方法,以关键词为条件语句得到的查询结果通常需要更多的二次处理来进行信息的过滤和筛选[1]。例如,读者期望通过检索关键词“BM25”来查阅 BM25 的技术细节以及算法改进,却返回大量关于将 BM25 应用于某一具体问题的文献。因此,标识学术文献中关键词的语义功能,能够为指向性的快速知识索引构建底层支持,对于知识精确检索和知识结构化表示均具有重要的理论意义和实践价值。

关键词语义功能的识别需在理解其上下文语境的基础上,充分挖掘潜在的写作规律来完成。然而,已有研究在词汇上下文建模中多局限于文字层面的符号语义表征,忽视了关键词位置及文本结构等重要信息,而这些特征在一定程度上能够侧面揭示关键词的功能角色。依据科技文献的行书范式,不同功能的关键词在全文中应具有不同的概率分布,文献[2]中作为研究问题的“图像分类”更可能会在引言或相关工作章节中进行着重描述,而作为研究方法的“支持向量机”则更倾向频繁显现于方法或实验章节。W. Lu 等[3]通过对文献关键词位置的统计分析发现关键词列表中的排列顺序也依循一定规律:描述问题或方法的关键词多处于列表的前置位,该现象在中文期刊论文中尤为突出。此外,由于不同学科所研究问题的差异,相同的关键词在不同学科或领域中存在不同的语义功能倾向。为有效表征并运用这些特征,本文分别设计了不同特征的表示方法,构建了一种多特征融合的关键词语义功能识别模型,通过引入关键词位置信息和先验知识信息,在充分捕获关键词上下文特征的基础上实现了关键词的语义功能识别。

* 本文系国家自然科学基金项目“基于多语义信息融合的学术文献引文推荐研究”(项目编号:71673211)和国家自然科学基金青年项目“基于深度语义挖掘的引文推荐多样化研究”(项目编号:71704137)研究成果之一。

作者简介:张国标(ORCID:0000-0002-1568-2492),博士研究生;李鹏程(ORCID:0000-0003-1427-7716),博士研究生;陆伟(ORCID:0000-0002-0929-7416),院长,教授,博士生导师,通讯作者,E-mail:weilu@whu.edu.cn;程齐凯(ORCID:0000-0003-3904-8901),副教授,博士。

收稿日期:2020-10-20 修回日期:2021-02-08 本文起止页码:89-96 本文责任编辑:杜杏叶

2 相关工作概述

关键词语义功能识别属于学术文本词汇功能研究的一部分,目前研究还处于初步探索阶段。早期的词汇功能研究主要围绕词汇的语法功能(Lexical Functional Grammar, LFG)[4]展开,应用基于统计的自然语言处理方法从语法结构层面分析词汇的主谓宾角色。其中,应用比较广泛的主要有条件随机场模型(CRF)和隐马尔可夫模型(HMM)。T. Moon等[5]通过分析文本内容和功能词之间的不同,设计了一种利用边界条件的HMM模型。孙静等[6]先用词典对文本进行词性标注,再利用CRF进行迭代标注,标注结果不断迭代优化。然而,词汇的语法功能仅能体现词汇之间的句法关系,无法从语义层面反映词汇的真实含义。T. Kondo等[7]从语义层面对词汇的功能进行了划分,定义了"领域""问题""方法"及"其他"4个词汇功能类别,并以此对特定领域中的技术路径和热点演化进行动态分析。随后,H. Nanba等[8]进一步对研究对象范围进行了扩展,对专利及科研文献的摘要进行词汇语义功能识别。S. Gupta等[9]借助句法模板以及重抽样在ACL论文数据集上进行了实验,结果显示在领域、问题、技术3个功能类别上的效果有了较大的提升,但依然无法达到实用水平。C. T. Tsai等[10]按技术、应用两类对词汇功能进行划分,并采用多特征结合和重采样的方法,取得了较好的性能表现。程齐凯等[11]对学术文献词汇功能的概念进行了进一步界定,设计了词法、句法、组块等27种特征,结合条件随机场构建了针对学术文献研究问题与研究方法的识别模型,并在GUPTA数据集上进行了验证。K. Hefernan等[12]为获取学术文献中的问题和方法,采用SVM等多种机器学习方法对文本中的词汇进行了问题方法的分类判别。陆伟等[13]以学术文献中的关键词作为语义功能识别对象,利用BERT及LSTM方法构建分类模型,对关键词所承载的问题和方法语义功能进行了分类。

总体而言,研究人员已对学术文本中的标题、摘要和关键词进行了词汇语义功能识别探索,在取得一些成果的同时也暴露出一些不足之处。由于词汇语义功能是指词汇在文本中承载的语义功能角色[11],已有方法仅针对词汇的上下文信息进行建模。然而,仅仅使用上下文依赖特征无法实现词汇语义信息的充分表征,词汇在文献中的位置信息与知识传承中的词汇使用习惯也在一定程度上揭示了其对应的语义功能。例如,在学术文献中问题类的词汇多出现于引言或文献回顾章节,方法类的词汇则更倾向于频繁出现在方法或实验章节。因此,本文拟结合关键词的位置信息和先验知识信息,构建关键词语义功能识别模型。

3 关键词多特征表示与融合

3.1 关键词多特征表示

3.1.1 上下文依赖特征

科技文献的标题是文章的眉目,是对全文的最精粹的概括,是对文章研究主题的揭示,往往能够体现文章的研究问题和研究的创新。科技文献摘要是对全文内容的精简陈述,具有短、精、完整三大特征,一般应说明研究目的、方法、结果和结论等。近年来科技期刊逐步要求作者提供结构式摘要,简明描述研究的主要内容。因此,可以使用标题和摘要信息来捕捉关键词的上下文依赖特征。在文献中提及问题和方法时,其上下文使用了较多的习惯语,如"基于/XX的/…""…采用/XX方法…"和"…是/XX问题/…"等,这些习惯性的写作模式构成了关键词上下文特征。本文采用目前文本处理效果最好的BERT模型来实现关键词上下文依赖特征的表示。

3.1.2 位置特征

不同语义功能类别的关键词,在文中的表述细节应具有一定区别。相对于其他复杂开放场景中的文本,学术文本通常具有严谨的逻辑结构和规范的层次,遵循科学研究的一般过程,从提出研究问题、介绍研究方法到结果的讨论和结论[11]。通常而言,一篇完整的研究性论文大致可分为5个章节块:引言、相关工作、方法、实验以及结论,各个章节在文中依次承担不同的结构功能。例如,引言的作用是对论文所涉及的背景、问题等进行初步的介绍,相关工作则是对问题相关文献进行系统的查阅分析。在这一机理下,不同语义功能类别的关键词,在不同章节中的表述细节应具有一定区别,描述研究问题的关键词会频繁出现于引言和相关工作章节,而描述研究方法的关键词则更倾向于表述在方法和实验章节。考虑到关键词在不同章节的词频信息能够在一定程度上揭示其相应的语义功能,本文将关键词的位置信息和词频信息进行向量化表示,通过统计关键词在各章节出现频次,来构造特征向量。

此外,对于关键词在关键词列表中的位置特征,本文采用One-hot编码形式对关键词列表排序信息进行表示,将关键词的位置序号处的数字设为1,其他位置数字设为0。

3.1.3 先验知识特征

相同的关键词在不同学科或领域中通常有着既定的功能倾向。例如,"支持向量机"在机器学习领域作为一种研究问题出现,通过改进以优化分类器的准确率和召回率;而在图像识别领域,"支持向量机"则更大概率作为研究方法出现。基于此规律即可以通过统计关键词的领域功能属性类别概率作为先验知识。本文同样采用概率形式对关键词在某一领域的问题方法属性进行表示,对于某一关键词,统计其在该领域数据集内作为问题出现的次数、作为方法出现的次数及作为其他功能出现的次数,计算概率,最终生成特征向量。

3.2 关键词多特征融合

多特征融合是将多种特征信息融合成一个新的特征向量用于下一步的词汇功能识别。不同特征的有效融合可以充分利用各特征所蕴含的类别信息以及充分发挥特征互补作用,然而简单的特征串联拼接没有充分考虑不同特征之间的差异性。针对这一问题,研究采用注意力机制(Attention)来区分不同特征在分类任务中的重要性,通过注意力概率分布突出特定特征对词汇功能识别任务的重要程度。Attention 函数的本质可以被描述为一个查询(query)到一系列键值(key-value)对的映射。在计算注意力概率时主要分为3步:①将 query 和每个 key 进行相似度计算得到权重;②使用 Softmax 函数对这些权重进行归一化;③将权重和相应的 value 进行加权求和得到最后的注意力概率,每步的计算方法如公式(1)－(3)所示:

$$L(H) = \tanh(wH + b) \qquad \text{公式(1)}$$

$$\alpha = soft\ \max(L(H)) = \frac{\exp(L(H))}{\sum_i \exp(L(H))_i} \qquad \text{公式(2)}$$

$$F = H \cdot \alpha \qquad \text{公式(3)}$$

公式(1)－(3)中,$L(H)$表示向量 H 对应特征的权重,w 表示权重系数,b 表示偏差,tanh 是激活函数,α 表示经过归一化后的特征权重,i 表示向量 H 的第 i 个值。

注意力机制对所构造的4种关键词特征进行加权变换,以突出重要的特征对词汇功能识别结果的贡献,提高模型的分类准确度。其流程见图1:针对输入的数据 X 通过特征表示输出所提取的4种特征,分别用 f1、f2、f3、f4 表示上下文依赖特征、关键词全文位置特征、关键词列表位置特征和先验知识特征;随后,对拼接后的关键词特征向量 $H = [f1,f2,f3,f4]$ 进行 tanh 函数权重计算,并通过 softmax 函数权重归一化得到其注意力概率 α。最后,将特征向量 H 与所得概率 α 进行点乘以实现加权融合,得到融合特征 F。

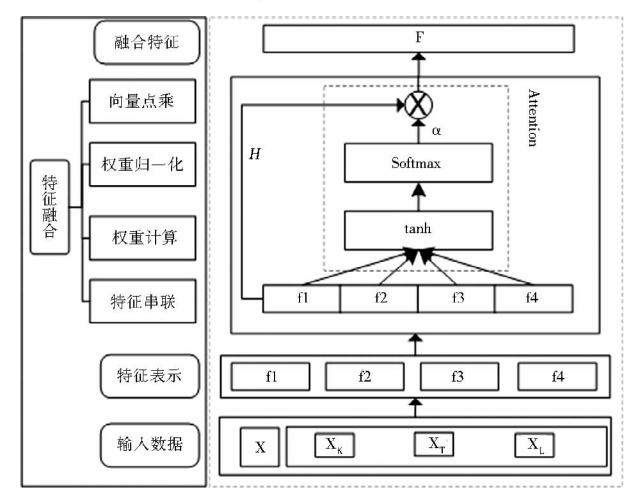


图1 关键词多特征融合方案

4 关键词语义功能识别模型构建

关键词语义功能识别的本质是信息抽取问题,鉴于深度学习方法在多类任务中的优异表现,文本拟将信息抽取问题转化为文本分类问题,通过采用深度学习可解的标签判定策略实现关键词的语义功能判别,构建语义功能识别模型。在对关键词的上下文依赖、位置以及先验知识等信息进行融合表征的基础上,进行基于深度学习模型的词汇功能识别模型设计,并以文本分类任务中表现最好的 BERT 模型作为上下文依赖表征模块,优化深度神经网络拟合非线性的能力,针对关键词有关研究问题和研究方法的语义功能进行识别。

4.1 词汇功能识别模型原型选择

在文本处理任务中,深度学习模型通常采取预训练词向量的方法为后续任务提供输入信号的向量化表示。词向量是通过浅层网络进行无监督训练将词汇映射到向量空间,如在文本处理中广泛使用的 Word2Vec[14] 模型,该方法通过构建词向量表将经过分词处理的文本映射到多维向量空间中。虽然该方法能够在词汇层面上具备较好的效果,但割裂式的向量化拼接使得语句级的向量化表示缺少对连续文本内在联系和语言结构的表达能力。除此之外,静态化的词嵌入方法在多义词上也存在着巨大的局限性,无法结合词汇的上下文语义进行动态向量化表示,如 apple 既可表示苹果也可指代 Apple 公司。针对现有方法的缺陷,Google AI 团队于2018年提出了一种基于 Transformer[15]

模型的预训练向量表示方法——BERT（Bidirectional Encoder Representation from Transformers）[16]。BERT 网络模型在遵循词嵌入一般思想的基础上，进一步增加了词向量模型的泛化能力，通过字符级、词汇级以及句子级的多粒度特征关系挖掘，力求能够对文本的词性、句法和语义等信息进行充分描述。由图 2 所示 BERT 向量构成可知，文本的最终向量化表示由词条嵌入（Token Embedding）、分割嵌入（Segment Embedding）、位置嵌入（Position Embedding）3 个部分拼接构成。相较传统词向量模型，BERT 网络模型选择深层双向的编码层完成词向量的学习。考虑到单词的语义依赖于其所在的上下文环境，即它左右两侧的某些词，采用结合前向和后向的双向 encoding 能够使得词向量具有上下文关联的能力，继而实现动态化的向量词义消歧。

Input	[CLS]	the	food	is	great	[SEP]	I	love	it	[SEP]
Token Embedding	$E_{[CLS]}$	$E_{[the]}$	$E_{[food]}$	$E_{[is]}$	$E_{[great]}$	$E_{[SEP]}$	$E_{[I]}$	$E_{[love]}$	$E_{[it]}$	$E_{[SEP]}$
	+	+	+	+	+	+	+	+	+	+
Segment Embedding	$E_{[A]}$	$E_{[A]}$	$E_{[A]}$	$E_{[A]}$	$E_{[A]}$	$E_{[A]}$	$E_{[B]}$	$E_{[B]}$	$E_{[B]}$	$E_{[B]}$
	+	+	+	+	+	+	+	+	+	+
Position Embedding	$E_{[0]}$	$E_{[1]}$	$E_{[2]}$	$E_{[3]}$	$E_{[4]}$	$E_{[5]}$	$E_{[6]}$	$E_{[7]}$	$E_{[8]}$	$E_{[9]}$

图 2　BERT 向量构成

BERT 模型综合运用词条嵌入、分割嵌入和位置嵌入 3 种信息，通过 Transformer 输出含有丰富语境信息的词向量，实现了不同语境下的词的动态向量化表示。在词汇功能识别问题中，同样需要根据关键词所在上下文语境实现关键词语义的动态表示，以在词汇功能识别模型训练过程中区分不同语境下关键词的不同语义功能。

4.2　多特征融合的关键词语义功能识别模型

针对准确理解关键词在科技文献中的语义功能的目的，本文融合关键词上下文依赖特征、位置特征和先验知识特征，在 BERT 模型的基础上构建关键词语义功能识别模型，在对多种特征进行向量化表示之后，将 4 种特征向量进行拼接，并采用注意力机制来实现不同特征的权重分配，最终采用 Softmax 分类器对融合向量进行分类，输出关键词功能类别。

依照上述设计路线，多特征融合的关键词语义功能识别模型可分为输入层、特征表示层、特征融合层、检测层，如图 3 所示：

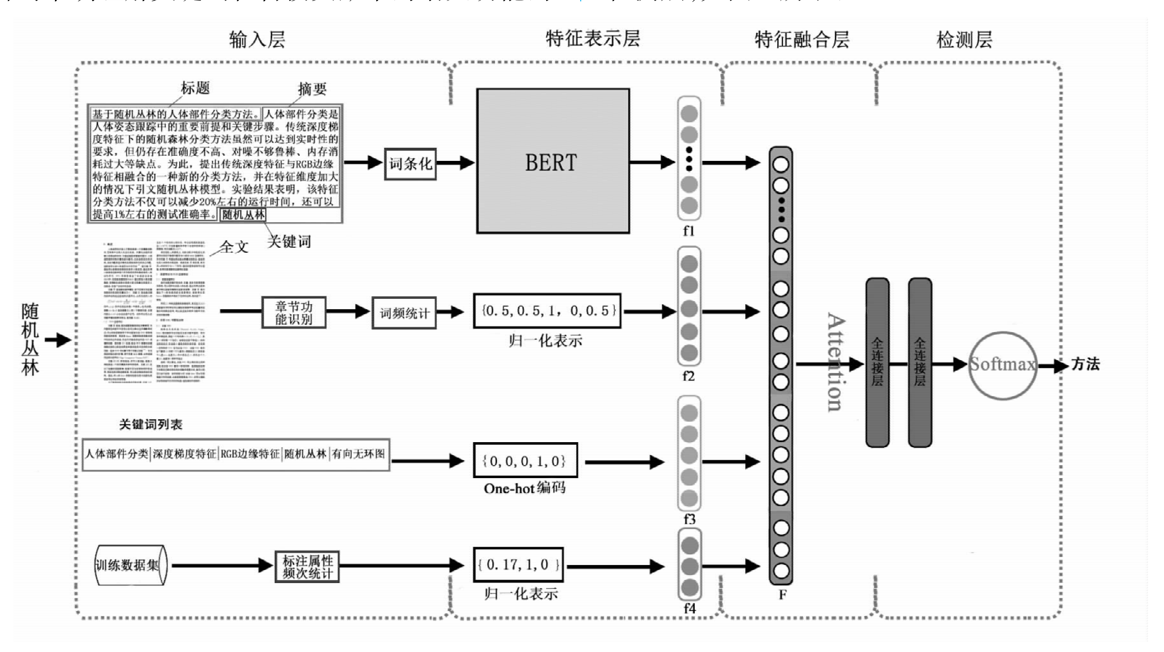


图 3　多特征融合的关键词语义功能识别模型

张国标，李鹏程，陆伟，等. 多特征融合的关键词语义功能识别研究[J]. 图书情报工作,2021,65(9):89-96.

输入层通过从原始数据中抽取能够输入到模型的信息,包括各特征对应的原始数据及对其所施加的预处理。对于上下文特征,本文通过将文章标题、摘要和关键词进行拼接来构成原始数据,通过词条化预处理,将文字序列分割开来,作为输入数据。对于全文位置信息,首先采用文献[17]所提出的章节功能识别方法,将全文各章节按照文献[18]提出的结构功能框架划分为引言、相关研究、研究方法、实验和研究结论 5 个章节,并统计关键词在各章节的出现频次,作为输入数据。对于关键词列表位置信息,通过匹配关键词列表获取关键词位置序号。对于先验知识,通过统计某一关键词在自建训练数据集中分别被标注为问题、方法、其他等 3 种类型功能的次数来获得关键词先验知识数据。

特征表示层针对预处理后的原始数据,采用不同方法实现原始数据的向量化表示。对于上下文依赖特征数据,采用 Google 预训练的中文 BERT 模型处理标题、摘要与关键词的拼接数据,将 BERT 模型输出的向量作为上下文依赖特征。对于全文位置统计数据,为方便数据处理通过归一化(Normalization)将位置频次转化为[0,1]之间的小数,生成 5 维特征向量。对于关键词列表位置序号数据,通过统计自建训练数据集文献的关键词数得出平均值 5,因此同样将关键词列表位置特征表示为 5 维向量,采用 One-hot 形式进行编码,并根据关键词位置序号将对应位置的数字设为 1,其他位置数字为 0。对于先验知识,通过统计某一关键词在自建训练数据集中分别被标注为问题、方法和其他 3 种类型功能的次数来获得关键词先验知识数据,并对其进行归一化,生成 3 维特征向量。

特征融合层采用 3.2 节所设计的融合方案,采用注意力机制将各种特征进行加权融合,实现各特征的权重分配。

输出层采用两层全连接网络作为检测模型,最后通过 Softmax 分类器对特征向量进行分类,输出结果标签。

5 关键词语义功能识别实验

5.1 数据标注

由于目前还没有标准的学术文献关键词语义功能识别语料库可用于实验研究,本文将根据自建的语料库完成对研究问题与研究方法的识别任务。自建数据集的数据来源为中国知网数据库收录的《计算机工程》《计算机科学》《计算机学报》《模式识别与人工智能》等期刊近 10 年(2009-2018 年)发表的 100 025 篇研究型论文的标题、摘要和关键词,根据论文标题和摘要对关键词进行了机器和人工标注。出于实用性考虑,采用一个简单通用的关键词语义功能分类方案,将计算机领域关键词语义功能界定为研究问题、研究方法和其他 3 个类别。对于规律性较强的标题,例如“基于 XX 的 XX”,采用模板匹配的方法进行数据标注,并进行人工审核。对于规律性不强的标题,采用人工标注,在标注过程中,共有 2 位情报学博士研究生和 2 位情报学硕士研究生参与数据标注工作,数据标注共经过两轮标注,第一轮由个人标注,第二轮标注对于个人无法确定的数据,由 4 位同学投票决定数据类型。经过人工标注和筛选,共获得 310 214 条数据,其中,标记为“方法”的关键词为 102 278 个,标记为“问题”的关键词为 102 504 个,标注为“其他”的关键词为 105 432 个。将 310 214 条标注数据按照 8:1:1 的比例划分为训练集、验证集和测试集,具体数量如表 1 所示:

表 1 实验数据统计

类别	训练集/条	验证集/条	测试集/条	总计/条
问题	82 003	10 250	10 251	102 504
方法	81 822	10 228	10 228	102 278
其他	84 346	10 543	10 543	105 432
总计	248 171	31 021	31 022	310 214

5.2 实验步骤与设置

实验在 Ubuntu16.4 操作系统 + Python3.6 编程环境下进行,采用 Tensorflow 深度学习框架构建关键词语义功能识别模型并进行训练。首先,按照关键词语义功能识别模型中设计的 4 种关键词特征进行特征抽取并保存,对于 f1 特征,利用 Google 提供的中文 BERT 预训练模型读取由标题、摘要、关键词的拼接而成的字符序列,设定最大字符长度为 512,并输出维数为 768 的向量;对于 f2 特征,由于无法获取文献全文数据,实验时采用摘要作为替代,设置摘要长度为 5 条语句,对于超过 5 条的截取前 5 条,对于不足 5 条语句的摘要进行空格填补,统计关键词在摘要每条语句中出现频次,并对统计结果进行归一化处理,输出 5 维向量;对于 f3 特征,根据关键词在关键词列表中出现的位置输出 4 维向量;对于 f4 特征,根据关键词在数据集中不同语义功能类型的出现频次进行归一化处理,输出 3 维向量。随后,根据实验需要对 4 种特征向量进行拼接组合,输入到由 Attention 机制和双全连接层构成的分类模型中,进行分类模型训练。为了防止过拟合和提高模型鲁棒性,采用指数衰减法优化深度学习速率(每训

练 500 步衰减 5%),在模型的全连接层添加了 Dropout,并在训练过程中采用了 Early stop 策略,并运用自动调参工具 Talos 进行了参数优化,最终参数如表 2 所示:

表 2　实验参数设置

参数	参数值
Epoch	10
Dropout	0.2
Batch_size	32
激活函数	Relu
学习率	0.002
全连接层神经元个数	400 100

5.3　实验结果与分析

实验采用文本分类常用的准确率(Accuracy)、精确率(Precision)、召回率(Recall)指标评价词汇功能识别模型的性能,并采用 F1(F1-measure)指标评价其综合性能。实验通过设置对照组分别检验不同特征模型的性能,在关键词语义功能识别实验中,实验目的为:

(1)验证 Attention 融合方法的有效性,对 4 种特征的拼接向量施加 Attention 权重获得融合向量,对比分析加 Attention 权重的和未加 Attention 权重的关键词各语义功能类别的宏平均结果。

(2)比较各类特征模型识别性能的影响,获取最有效的特征组合。首先计算只采用 BERT 上下文依赖特征时的模型性能,并以此为基础,采用 Attention 融合方案分别计算不同特征与 BERT 上下文依赖特征组合时模型在各类语义功能上表现性能的宏平均结果,以确定各特征融合后的效果差异。

(3)比较不同类型语义功能识别效果,分析不同类别语义功能识别差异。选取步骤 1 中识别效果最好的特征组合,分别比较模型对方法、问题和其他 3 种类型词汇的识别性能。

实验结果如表 3－表 5 所示:

表 3　特征融合方案性能对比

特征融合方案	Recall	Precision	F1	Accuracy
拼接 + Attention	0.973	0.973	0.973	0.978
拼接	0.965	0.967	0.965	0.972

从实验结果表 3 中可以看出,采用拼接 + Attention 的融合方案各评价指标均高于仅采用拼接进行融合的方案。这一结果说明,Attention 机制能够有效区分特征向量中各特征值的权重,对重要特征值赋予更高的权重,为后续处理传递关键及重要信息,使模型做出更加准确的判断。然而,此处的 Attention 机制对拼接后的特征向量进行加权,无法区分 4 种特征的特征权重,仅是对拼接后的特征值进行加权。

表 4　不同特征组合模型识别结果对比

特征	Recall	Precision	F1	Accuracy
f1	0.917	0.920	0.918	0.924
f1 + f2	0.920	0.930	0.925	0.933
f1 + f3	0.923	0.923	0.923	0.930
f1 + f4	0.967	0.967	0.967	0.974
f1 + f2 + f3	0.917	0.937	0.927	0.938
f1 + f3 + f4	0.967	0.970	0.968	0.975
f1 + f2 + f4	0.973	0.963	0.968	0.970
f1 + f2 + f3 + f4	0.973	0.973	0.973	0.978

表 5　所有特征融合的词汇功能识别结果对比

类别	Recall	Precision	F1
方法	0.980	0.980	0.980
问题	0.960	0.970	0.965
其他	0.980	0.970	0.975

从实验结果表 4 中可以发现,各种特征的加入均有效提升了 F1 值,融合所有特征的模型取得了最优效果,召回率为 0.973,精确率为 0.973,F1 值为 0.973,准确率为 0.978。这一结果说明所设计的关键词特征能够有效提升关键词语义功能识别性能。通过对比加入了先验知识特征和没加入先验知识特征的模型性能,可以发现加入先验知识特征之后模型分类效果有了较大提升,说明实验数据中的论文研究多遵循先验知识规律,对于模型识别效果有着较大帮助。通过对比关键词位置特征实验结果发现,关键词位置特征对于模型的精确率有一定的提升,而召回率提升有限,说明关键词位置特征能够帮助区分关键词功能类别,而对关键词类别内的规律挖掘有限。

从实验结果表 5 中可以看出,研究方法的精确率为 0.98,召回率为 0.98,F1 值为 0.98,在各个指标上均高于或等于其他类别,这是由于研究方法的创新难度较大,大部分研究均使用已有的研究方法。特定领域中研究方法,通常均能找到所对应的约定表述方式,模型能够更好地对其进行识别和判定。而对于问题类关键词,在研究过程中较为容易发现新的研究问题,其表达形式则显得更为多变和复杂,故其 F1 值和准确率相对较低。

6　结语

研究学术文献中的关键词语义功能可以从更深层次或者偏重于语义的角度理解学术文本词汇级别的结

构,以便理解不同词汇在文献中的功能角色,进而提升学术文献检索系统的准确性,缩短用户搜寻时间。本文在关键词上下文依赖特征的基础上,结合关键词位置信息和先验知识对关键词语义功能进行了识别,设计了关键词位置特征和先验知识特征表示方法,并采用深度学习方法构建了关键词语义功能识别模型。在自建计算机期刊文献数据集上的实验表明,本文所构建的多特征融合的关键词语义功能识别模型识别平均精确率达到了 0.973,优于只采用部分特征的方法。

由于词汇功能标注数据集匮乏,本文仅针对学术文献中的关键词进行了语义功能识别,且仅识别了问题与方法两种语义功能,未能实现学术文本全文词汇的多类型语义功能识别。未来将不断积累词汇语义功能数据,构建更大规模的数据集,并设计更为全面的词汇语义功能类型。同时,基于对学术文本全文的认知理解,深入挖掘潜在论文写作规律,探索词汇功能识别新方法,以期找到最有效的学术文本词汇功能识别方法,并使之能够应用于其他的学术文本挖掘研究中。另外,在完成学术文本词汇功能识别之后,如何进一步实践应用同样是下一步研究的重点,在未来的研究中可将词汇功能应用于文献推荐、关键词抽取、自动摘要和知识图谱等研究中,不断拓展词汇功能应用领域。

参考文献:

[1] 杨涛. 中文智能搜索引擎浅析[J]. 图书情报工作,2002,56(1):62－65.
[2] 万华林, CHOWDHURY M U. 基于支持向量机的图像语义分类[J]. 软件学报,2003,14(11):1891－1899.
[3] LU W, LI X, LIU Z F, et al. How do author-selected keywords function semantically in scientific manuscripts[J]. Knowledge organization,2019,46(6):403－418.
[4] DALRYMPLE M, KAPLAN R M, MAXWELL J T, et al. Formal issues in lexical-functional grammar[M]. Stanford: CSLI Publications, 1995.
[5] MOON T, ERK K, BALDRIDGE J. Crouching dirichlet, hidden markov model: unsupervised POS tagging with context local tag generation[C]//Proceeding of the 2010 conference on empirical methods in natural language processing. Cambridge:Association for Computational Linguistics, 2010:196－206.
[6] 孙静,李军辉,周国栋. 基于条件随机场的无监督中文词性标注[J]. 计算机应用于软件,2011,28(4):21－23.
[7] KONDO T, NANBA H, TAKEZAWA T, et al. Technical trend analysis by analyzing research papers' titles[M]. Berlin:Springer, 2009.
[8] NANBA H, KONDO T, TAKEZAWA T. Automatic creation of a technical trend map from research papers and patents[C]//Proceedings of the 3rd international workshop on patent information retrieval. New York: Association for Computing Machinery,2010:11－16.
[9] GUPTA S, MANNING C. Analyzing the dynamics of research by extracting key aspects of scientific papers[C]//Proceedings of 5th international joint conference on natural language processing. Stroudsburg:Association for Computational Linguistics, 2011:1－9.
[10] TSAI C T, KUNDU G, DAN R. Concept-based analysis of scientific literature[C]//Proceeding of the 22nd acm international conference on conference on information & knowledge management. New York:Association for Computing Machinery,2013:1733－1738.
[11] 程齐凯,李信. 面向语义出版的学术文本词汇语义功能自动识别[J]. 数字图书馆论坛, 2017,159(8): 24－31.
[12] HEFFERNAN K, TEUFEL S. Identifying problems and solutions in scientific text[J]. Scientometrics,2018, 116(1):1－16.
[13] 陆伟,李鹏程,张国标,等. 学术文本词汇功能识别——基于BERT 向量化表示的关键词自动分类研究[J]. 情报学报, 2020,39(12):1320－1329.
[14] 唐明,朱磊,邹显春. 基于 Word2Vec 的一种文档向量表示[J]. 计算机科学,2016,43(6):214－217,269.
[15] VASWANI A, SHAZEER N, PARMAR N. Attention is all you need[C]//Proceedings of the 31th international conference on neural information processing systems. New York:Curran Associates Inc, 2017:6000－6010.
[16] DEVLIN J, CHANG M W, LEE K, et al. Bert: pre-training of deep bidirectonal transformers for language understanding[C]//Proceeding of the 2019 conference of the North American chapter of the Association for Computational Linguistics:human language technologies. Minneapolis: Association for Computational Linguistics, 2019:4171－4186.
[17] 王佳敏,陆伟,刘家伟,等. 多层次融合的学术文本结构功能识别研究[J]. 图书情报工作, 2019,63(13):95－104.
[18] 黄永,陆伟,程齐凯,等. 学术文本的结构功能识别——基于段落的识别[J]. 情报学报, 2016,35(5):530－538.

作者贡献说明:

张国标:论文撰写,数据标注和实验分析;
李鹏程:数据标注,实验方案设计;
陆伟:研究思路和框架提出,论文修改;
程齐凯:实验方案设计,论文修改。

Research on Keyword Semantic Function Recognition Based on Multi-feature Fusion

Zhang Guobiao[1,2] Li Pengcheng[1,2] Lu Wei[1,2] Cheng Qikai[1,2]
[1] School of Information Management, Wuhan University, Wuhan 430072
[2] Institute for Information Retrieval and Knowledge Mining, Wuhan University, Wuhan 430072

Abstract: [Purpose/significance] Keywords, as a kind of vocabulary or term that can reveal the subject and core content of a text, can identify the functions and provide the underlying index support for fast and accurate acquisition of knowledge and documents. [Method/process] Aiming at the existing studies that are mostly limited to the semantic representation of symbols at the text level in vocabulary context modeling, this paper proposes a lexical function recognition model based on multi-feature fusion. On the basis of capturing the context-dependent features of keywords using the BERT model, the position information of keywords in the keyword list and the full text and prior knowledge of vocabulary functions are fused, and then the attention mechanism and feed-forward neural network are used for the identification of key words by problem-solving method. [Result/conclusion] The experimental results show that both the location information and priori knowledge of the keywords can improve their word function recognition effect, and the prior knowledge has a greater contribution to the recognition effect.

Keywords: lexical function recognition academic text keyword BERT multi-feature fusion